Efficient selection of branch-specific models of sequence evolution.

Julien Y Dutheil, Nicolas Galtier, Jonathan Romiguier, Emmanuel Douzery, Vincent Ranwez, Bastien Boussau

Efficient Selection of Branch-Specific Models of Sequence Evolution

Julien Y. Dutheil,*[,1] Nicolas Galtier,[1] Jonathan Romiguier,[1] Emmanuel J.P. Douzery,[1] Vincent Ranwez,[1,2] and Bastien Boussau[3,4]

[1]Institut des Sciences de l'Évolution—Montpellier, Université Montpellier 2, Montpellier, France
[2]Montpellier SupAgro, UMR AGAP, Montpellier, France
[3]Laboratoire de Biométrie et Biologie Évolutive, Université de Lyon, Université Lyon 1, CNRS, UMR5558, Villeurbanne, France
[4]Department of Integrative Biology, University of California Berkeley

***Corresponding author:** E-mail: julien.dutheil@univ-montp2.fr.

Associate editor: Jeffrey Throne

Abstract

The analysis of extant sequences shows that molecular evolution has been heterogeneous through time and among lineages. However, for a given sequence alignment, it is often difficult to uncover what factors caused this heterogeneity. In fact, identifying and characterizing heterogeneous patterns of molecular evolution along a phylogenetic tree is very challenging, for lack of appropriate methods. Users either have to a priori define groups of branches along which they believe molecular evolution has been similar or have to allow each branch to have its own pattern of molecular evolution. The first approach assumes prior knowledge that is seldom available, and the second requires estimating an unreasonably large number of parameters. Here we propose a convenient and reliable approach where branches get clustered by their pattern of molecular evolution alone, with no need for prior knowledge about the data set under study. Model selection is achieved in a statistical framework and therefore avoids overparameterization. We rely on substitution mapping for efficiency and present two clustering approaches, depending on whether or not we expect neighbouring branches to share more similar patterns of sequence evolution than distant branches. We validate our method on simulations and test it on four previously published data sets. We find that our method correctly groups branches sharing similar equilibrium GC contents in a data set of ribosomal RNAs and recovers expected footprints of selection through dN/dS. Importantly, it also uncovers a new pattern of relaxed selection in a phylogeny of Mantellid frogs, which we are able to correlate to life-history traits. This shows that our programs should be very useful to study patterns of molecular evolution and reveal new correlations between sequence and species evolution. Our programs can run on DNA, RNA, codon, or amino acid sequences with a large set of possible models of substitutions and are available at http://biopp.univ-montp2.fr/forge/testnh.

Key words: molecular phylogenetics, maximum likelihood, ancestral character reconstruction, dN/dS, paml, selection.

Introduction

Living organisms show a striking diversity in size, life history, ecology, population structure, physiology, and cellular biology. This diversity propagates to the genome level: substantial between-species variations in base or amino acid compositions and in rates of sequence evolution have been documented (Hickey and Singer 2004; Bromham 2009; Lartillot and Poujol 2011). Understanding the links between these two kinds of variations—phenotypic versus genomic—is an important goal of current molecular evolutionary research (Boussau and Daubin 2010).

A popular approach to this question is to correlate phenotypic and sequence evolution across the branches of a phylogenetic tree (Yang 1998; Paland and Lynch 2006; Boussau et al. 2008). This typically requires identifying groups of branches (e.g., subtrees) sharing a common molecular evolutionary process. Usually this branch-clustering step is performed a priori and reflects existing knowledge (or hypothesis) about the major factors affecting sequence evolution in the group of organisms under study.

For instance, many studies of lineage-specific variations in selective pressure rely on PAML (Yang 2007) and use one of two distinct procedures: 1) the user either a priori defines clusters of branches that are expected to show a similar ratio of nonsynonymous to synonymous codon substitutions (dN/dS) or 2) assumes that each branch has an idiosyncratic dN/dS. The second option is discouraged by the PAML manual as it requires estimating one parameter per branch of the tree, which tends to be unstable when the tree is large. The first approach can only be applied in cases where prior knowledge, for instance on the phenotype of organisms, is available to cluster branches. Even in these few cases where phenotypes of extant organisms are well known and can be assumed to drive sequence evolution, it is often difficult to determine a priori how internal branches should be clustered as they correspond to organisms whose phenotypes can no longer be observed. An alternative approach is therefore desirable, which would avoid overparameterization and arbitrary grouping of branches prior to sequence analysis. With the increasing facility for gathering sequence data in living organisms of any sort, the tree-partitioning issue has lately become prominent in the molecular evolutionary literature and has motivated specific methodological developments (Jayaswal et al. 2011, Zhang et al. 2011).

We present a statistical approach to cluster branches in a phylogenetic tree using only sequence information. This approach aims at identifying the optimal partition of the set of branches in a likelihood framework according to Akaike information criterion (AIC) or Bayesian information criterion (BIC) and by design avoids overparametrization and subjective decisions from the user. The objective is to group branches along which sequence evolution has been similar, in terms of any set of descriptive statistics of the substitution matrix. These statistics can be, for instance, dN/dS, equilibrium GC content, or the ratio of conservative to nonconservative amino acid substitutions (Sainudiin et al. 2005). The partition of branches returned by our approach can be correlated to characteristics of the organisms under study in order to identify new links between phenotypic and genomic features and possibly deduce ancestral characters at internal nodes of the tree. Such an approach is similar in principle to the so-called local molecular clocks, where branches with similar rates are clustered (Yang and Yoder 2003; Aris-Brosou 2007; Drummond and Suchard 2010; Heath et al. 2011). In our case, however, we are interested in the differential rate of each type of substitution, not the global amount of substitutions.

In this manuscript, we first present a new heuristic algorithm to find the optimal partition of branches and estimate parameters of substitution matrices. This algorithm benefits from an initial step of substitution mapping (Nielsen 2002; Rodrigue et al. 2008; Minin and Suchard 2008) and is implemented in C++ using the Bio++ libraries (Dutheil et al. 2006). Then, using simulations, we show that our approach is both fast and accurate. We apply our methods to previously published data sets, compare it with other approaches, and demonstrate that our new algorithms allow one to reveal the phenotypic determinants of sequence evolution for a wide range of experimental conditions, including large data sets.

Materials and Methods

Definitions and Notations

We denote D_i the ith site of the data set, that is, a column of the alignment, and Θ the set of parameters of a given model of sequence evolution. We consider the tree topology as fixed. By convention, we consider top nodes to be closer to the leaves than bottom nodes.

Substitution Mapping

We count the number of substitutions that occurred on each branch of a phylogenetic tree and at each site in a sequence alignment. We extend the procedure described in Dutheil et al. (2005) by providing detailed counts for each type of substitution in lieu of the total number of substitutions, following work by Minin and Suchard (2008) and Hobolth and Stone (2009).

We recall that at each position i in the alignment, we can compute the substitution vector $V_i^s = (\nu_{i,1}^s, \ldots, \nu_{i,b}^s, \ldots, \nu_{i,m}^s)$, where $\nu_{i,b}^s$ is the posterior estimate of the number of substitutions of type s that occurred on branch b and m is the number of branches in the tree. Following Dutheil et al. (2005), $\nu_{i,b}^s$ is estimated by averaging over all possible ancestral states at top (x_q) and bottom (x_p) nodes of branch b:

$$\nu_{i,b}^s = \sum_{x_p} \sum_{x_q} \Pr(x_p, x_q | D_i, \Theta) \times n_{x_p,x_q}^s(t). \quad (1)$$

In this equation, $\Pr(x_p, x_q | D_i, \Theta)$ is the joint probability of having state x_p at bottom node and state x_q at top node given the data and parameters. It is computed as follows (Galtier and Boursot 2000; Pupko et al. 2003; Dutheil et al. 2005):

$$\Pr(x_p, x_q | D_i, \Theta) = \frac{\Pr(x_p, x_q, D_i | \Theta)}{\Pr(D_i | \Theta)}. \quad (2)$$

The denominator is the likelihood for site i (Felsenstein 1981), whereas the numerator is obtained in a very similar way but considering the ancestral states x_p and x_q as known in the Felsenstein recursion. Term $n_{x_p,x_q}^s(t)$ is the mean number of substitutions of type s that occurred on a branch of length t knowing initial state x_p and final state x_q. Several methods have been proposed to compute this mean number. Instead of the method of Dutheil et al. (2005), we use the uniformization method proposed by Hobolth and Stone (2009) and Tataru and Hobolth (2011) because it is exact, numerically more stable, and for each site and branch it returns an array containing counts for each type of substitution. For alphabets with high dimension like the codon alphabets, substitution types can be summed, for example, all synonymous substitutions and all nonsynonymous substitutions, generating two types of counts instead of $61 \times (61 - 1) = 3,660$. As shown in Dutheil et al. (2005), these equations can be easily extended to account for variation of the substitution rate across sites and do not depend on the model used to describe the rate variation. For codon models, a constant distribution of codon substitution rates was used, as in the PAML software, whereas for nucleotide sequences we used a gamma distribution of site-specific substitution rates.

We summed substitution counts obtained for each site of the alignment to obtain branch-wise counts for each type of substitution. We further pooled substitution counts depending on the biological question we addressed: A or $T \rightarrow G$ or C and G or $C \rightarrow A$ or T in order to study the variation of GC content in nucleotide sequence and synonymous versus nonsynonymous substitutions for studying the variation of selection regime in codon sequences. Had we chosen to use our method to study selection at the amino acid level, we could have pooled substitutions in conservative versus nonconservative substitutions.

The substitution mapping requires a model of sequence evolution and a phylogenetic tree to work with. Several works have shown that this procedure is robust to the input model of sequence evolution (see, for instance, Minin and Suchard 2008). We therefore fitted a Tamura (1992) (respectively Nielsen and Yang 1998) homogeneous model for the ribosomal RNA (rRNA) (respectively codon) data set and estimated all numerical parameters, that is, kappa, theta

(respectively omega), and branch lengths. These parameters were then used for mapping substitutions. For the rRNA data set, substitution mapping was performed on a rooted tree as nonstationary models will be used in the model selection procedure.

Measuring Substitution Process Homogeneity between Branches

The total numbers of substitutions of each type, ν_b^s, were computed for each branch by summing over all sites:

$$\nu_b^s = \sum_i \nu_{i,b}^s. \quad (3)$$

We designed a multinomial likelihood-based measure of substitution process homogeneity between any two branches, here named b_1 and b_2. Under the null model of homogeneous process, the two vectors of counts are assumed to be drawn from a unique multinomial distribution in which the probabilities of each type of substitution, p_s, are shared by the two branches. The likelihood of the set of counts under the null model is

$$L_0 = \frac{(\nu_{b_1} + \nu_{b_2})!}{\prod_s (\nu_{b_1}^s + \nu_{b_2}^s)!} \prod_s p_s^{(\nu_{b_1}^s + \nu_{b_2}^s)}, \quad (4)$$

where $\nu_{b_1} = \sum_s \nu_{b_1}^s$ is the total number of substitutions summed across categories $(\text{and similarly for } \nu_{b_2})$.

The maximum likelihood estimates of the p_s's, to be used in equation (4), are

$$p_s = \frac{\nu_{b_1}^s + \nu_{b_2}^s}{\nu_{b_1} + \nu_{b_2}}. \quad (5)$$

Under the alternative model of heterogeneous process, the two vectors of counts are drawn from two distinct multinomial distributions, the probabilities of substitution types being different between branches. The likelihood of the data is now

$$L_1 = \frac{\nu_{b_1}!}{\prod_s \nu_{b_1}^s!} \prod_s p_{1s}^{\nu_{b_1}^s} \times \frac{\nu_{b_2}!}{\prod_s \nu_{b_2}^s!} \prod_s p_{2s}^{\nu_{b_2}^s} \quad (6)$$

and the maximum likelihood estimates of the p_{1s}'s and p_{2s}'s are

$$p_{1s} = \frac{\nu_{b_1}^s}{\nu_{b_1}}, \quad p_{2s} = \frac{\nu_{b_2}^s}{\nu_{b_2}}. \quad (7)$$

Twice the log-likelihood ratio between the two models is calculated and compared with a chi-squared distribution, the number of degrees of freedom being equal to the number of categories minus one. The resulting P value is considered as a measure of compatibility between the two considered branches. This measure accounts for the uncertainty due to stochastic errors in substitution counts.

This method compares substitution counts between branches. Substitution counts, however, are not a full description of the substitution process when sequences are not at compositional equilibrium. Consider, for instance, two branches in each of which exactly 10 $AT \rightarrow GC$ and 10 $GC \rightarrow AT$ changes were counted. Now suppose that the GC content of the considered sequence is 90% for branch 1 and 10% for branch 2. Despite having identical substitution counts, these two branches have distinctive evolutionary processes: the per AT site $AT \rightarrow GC$ substitution rate, for instance, is nine times higher in branch 1 than in branch 2. To account for this effect, we define corrected counts $\nu_b'^s$ as

$$\nu_b'^s = \Lambda \frac{\nu_b^s}{k_b^s}, \quad (8)$$

where k_b^s is the number of positions in the sequence in branch b at which a substitution of category s could have occurred (e.g., for $AT \rightarrow GC$ substitutions, the number of A and T positions). The k_b^s's are estimated by reconstructing the distribution of ancestral sequences at the parent node of the considered branch (Yang and Roberts 1995). In equation (8), $\Lambda = \sum \nu_b^s / \sum \frac{\nu_b^s}{k_b^s}$ is a scaling factor ensuring that the sum (over substitution categories) of the $\nu_b'^s$'s is equal to the sum of the ν_b^s's. In this study, the exact counts ν_b^s were used in equations (4)–(7) for comparisons between synonymous and nonsynonymous substitutions and the corrected counts $\nu_b'^s$ for comparisons between $AT \rightarrow GC$ and $GC \rightarrow AT$ substitutions.

Partitioning Branches

We developed a new hierarchical clustering procedure in order to define subsets of branches along the phylogenetic tree, based on their respective substitution processes, as inferred from branch-wise counts for each type of substitution. The procedure clusters branches of the tree following a neighbor-joining strategy. Each branch is initially assigned its own cluster, then the two most similar clusters of branches are repeatedly merged until only one cluster is left. The resulting tree is assigned branch lengths so that the height of inner nodes reflect the P value associated to the underlying clusters. More precisely:

Initialization. We start by associating each branch to their own subset. All pairs of single-branch subsets are tested for homogeneity of substitution process using the previously introduced branch–pair homogeneity test. The corresponding P values are stored.

Extension. The two subsets with the highest P value, $P_{\max}$, are gathered into a new subset. A new node in the clustering tree is created, with height equal to $\frac{(1-P_{\max})}{2}$. Substitution counts for the new subset are obtained by summing the counts for each individual subset. The new subset is tested against all other subsets using the multinomial test and the summed counts.

Termination. The procedure stops when there is only one cluster left.

The resulting clustering tree defines a hierarchy of nonhomogeneous models, from one substitution matrix per branch of the phylogenetic tree to one substitution matrix only for the whole phylogenetic tree. These models can be optimized in the maximum likelihood framework and compared using AIC or BIC (see below).

We introduced two modifications to improve the robustness of the above clustering algorithm to short branches. First, branches for which no substitutions (or by extension, for which less than a user-specified number of substitutions) are inferred are automatically clustered with their parent node in the phylogenetic tree. Second, we treat negative branch lengths in the clustering tree as equal to zero and the corresponding parent node as multifurcating, therefore defining more than one new cluster of branches compared with the immediately simpler model.

Finally, we introduced a variant of the clustering algorithm that only allows neighbor branches to be clustered. This was obtained by modifying the initialization step so that non-neighbor branches have a negative *P* value, which prevents them from being clustered at any step. Similarly, during the extension step, non-neighbor subsets are assigned a negative *P* value. We refer to this variant as the "join" model, whereas we refer to the original clustering without neighbor constraint as the "free" model.

Model Selection

Models are constructed from a set of branch clusters and optimized in the maximum likelihood framework. Models built from more than one subset are nonhomogeneous and are constructed as described in Dutheil and Boussau (2008). Here we consider nonhomogeneous models derived from a single type of substitution matrix (e.g., Tamura 1992; Nielsen and Yang 1998), and only branch lengths and parameters of these substitution matrices are allowed to vary on a per-branch basis (θ for Tamura 1992, ω for Nielsen and Yang 1998). Despite these restrictions, such models encompass the vast majority of nonhomogeneous models used in the literature.

Nested models are obtained by successively considering each node of the clustering tree in order of increasing height, starting from the root. The root node defines a bipartition (two subsets) of branches, and each descending node iteratively splits one of the previously defined subsets of branches. Nested models are compared on the basis of their maximum likelihood. We implemented two comparison procedures: AIC and BIC. Model exploration is stopped when the scores of *n* consecutive models are lower than the current best score, where *n* is a positive number defined by the user, or when all models have been tested.

Note that in the case of nonstationary models like the Galtier and Gouy (1998) nucleotide model, we tested both the homogeneous and the homogeneous nonstationary models, which has one extra parameter, the root equilibrium GC frequency.

Simulations

In order to assess the performance of the method (clustering + model selection), we conducted a simulation analysis under various models. We used a random tree containing 25 leaves, generated under a Yule distribution with the *rtree* command from the R package APE (Paradis et al. 2004). We randomly generated 18 nonhomogeneous models according to the free setting by splitting the tree in 1, 2, 3, 4, 5, or 10 subsets of branches, with three replicates in each case. Each branch was assigned one of the partitions randomly. We also generated 18 nonhomogeneous models according to the join setting, by picking a random subtree and assigning it to a partition number. We simulated codon sequences under the YN98 model of sequence evolution in which the omega (=dN/dS) parameter of each subset of branch was drawn from a gamma distribution with $\alpha = \beta = 0.5$. Six sequence alignments were generated for each of the 18 models, three with 300 positions and three with 1,000 positions, totalling 108 sequence alignments. On each data set, we fitted a YN98 homogenous model with the true phylogeny and used it to perform substitution mapping. The resulting substitution maps were used to obtain a tree of clustered branches which in turn was used to guide model selection using BIC. We used the free (respectively join) clustering algorithm on the data sets simulated under free (respectively join) model. Three additional nested models were tested after a local minimum was found, and the resulting global minimum used for creating partitions.

Programs and Examples

The testnh package is available at http://biopp.univ-montp2.fr/forge/testnh as source code and binary versions. The programs are written in standard C++ and compile on Linux, Windows, and MacOS systems and only depend on the Bio++ libraries (Dutheil et al. 2006). The package contains two major programs: mapnh for performing the substitution mapping and clustering of branches and partnh for fitting substitution models on the resulting subsets. All the data sets analyzed in this article are provided as example files in the source distribution. The package also contains the randnh program that was used to generate random models in the simulation analysis.

Data Sets

Tree of Life Concatenated rRNA *Alignment*

We used the rRNA alignment (concatenated small and large subunits) and corresponding phylogeny from Boussau and Gouy (2006) previously analyzed in Dutheil and Boussau (2008). The tree was midpoint-rooted. The alignment contains 527 complete sites for 92 sequences including 22 archaea, 34 bacteria, and 36 eukaryotes. These species come from a large variety of environments and show a wide range of optimal growth temperatures, which is known to affect rRNA GC content evolution in prokaryotes (Boussau and Gouy 2006).

Lysozyme Data Set

This is the data set provided together with the PAML software and described in Yang (1998). This data set was also analyzed by Zhang et al. (2011). We used the tree shown in figure 2 in Zhang et al. (2011) where the human branch was removed in order to compare results.

Daphnia Data Set

Accession numbers given in Paland and Lynch (2006) were used to download mitochondrial sequences from 28

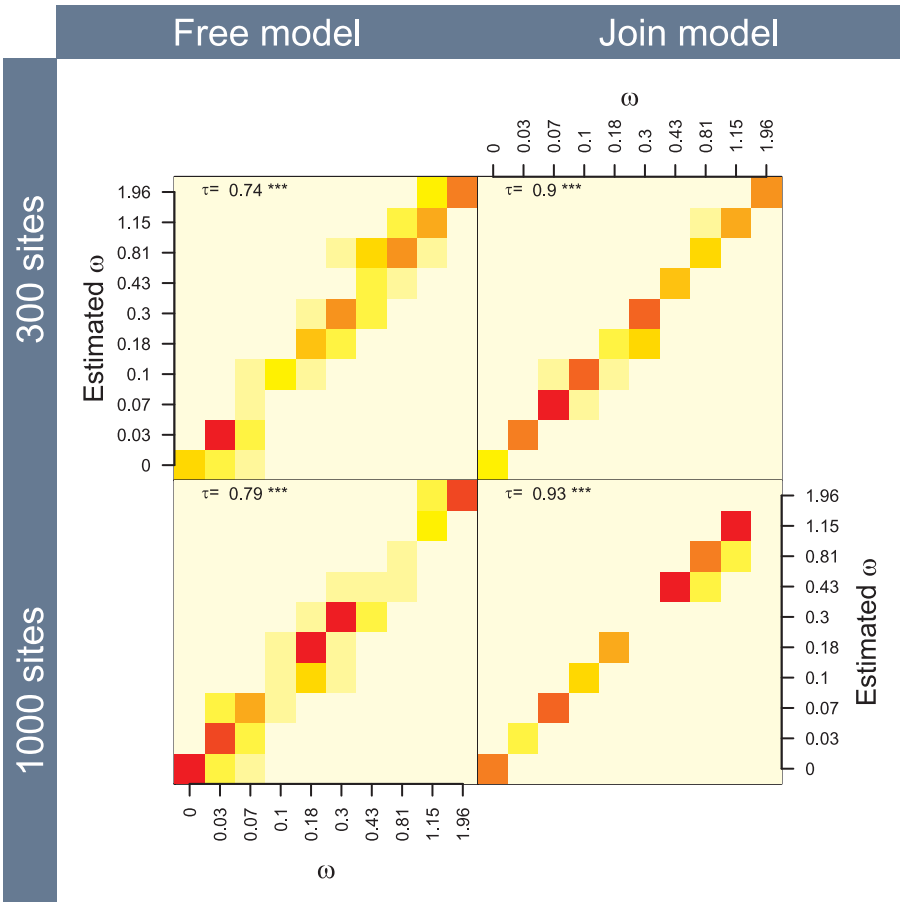

FIG. 1. Recovery of the branch-wise dN/dS (ω), displayed as a heat map. The *x* axis represents the true value of ω as used in the simulations, and the *y* axis displays the corresponding value estimated by the selected model. Colors describe the density of points, as 12 equispaced categories, from white (no point) to red (maximum number of points). Simulations with 1, 2, 3, 4, 5, and 10 random subsets are used (see Materials and Methods and supplementary fig. 1 for separate plotting). Values of the Kendall correlation test are reported for each model.

Daphnia pulex strains from GenBank (accession numbers DQ340817–DQ340843 and AF117817). Coding DNA sequences were extracted for genes *ATP6*, *ATP8*, *CO1*, *CO2*, *CO3*, *CYTB*, *ND1*, *ND2*, *ND3*, *ND4*, *ND4L*, *ND5*, and *ND6*, and concatenated. Alignments were done with Muscle (Edgar 2004) on amino acid sequences using Seaview (Gouy et al. 2010) and corrected by eye where necessary. This resulted in 3,681 codon sites.

Mantellid Frogs Data Set

The alignment and tree used in our analyses are as in Boussau et al. (2011) and can be downloaded from Treebase (http://purl.org/phylo/treebase/phylows/study/TB2:S11392).

Results

Simulations

Using simulated data with a codon model, we assessed the ability of our branch-clustering method to recover branch-specific parameter estimates. Figure 1 shows that the dN/dS values estimated by our clustering approach are accurately estimated, both in the free and in the join models. The join approach, however, shows more accurate estimates than the free approach. This was expected, as for the same number of partitions, the free model has more breakpoints (i.e. change in the substitution process) along the phylogeny than the join model, making it more difficult to recover for a data set of equivalent size. The larger data sets (1,000 sites) expectedly display a lower dispersion of estimates than the smaller ones (300 sites). The dispersion also increases with the number of partitions used in simulations (see Supplementary Material online), and the effect is stronger on the free model.

We also used these simulations to assess whether the number of clusters returned by the methods are meaningful. This is more complicated to evaluate as the number of clusters depends on the criterion used for model selection and the size of the data set as larger data sets have more power to discriminate small changes in the substitution process along the tree. Figure 2 displays the number of partitions recovered with the BIC criterion. The join approach performs slightly better than the *free* approach, which can again be explained by the fact that join data sets display less breakpoints than the free ones, given a certain number of partitions. Also apparent is that larger data sets (1,000 sites) tend to recover more clusters than smaller data sets (300 sites). In our simulation setup, this leads to an

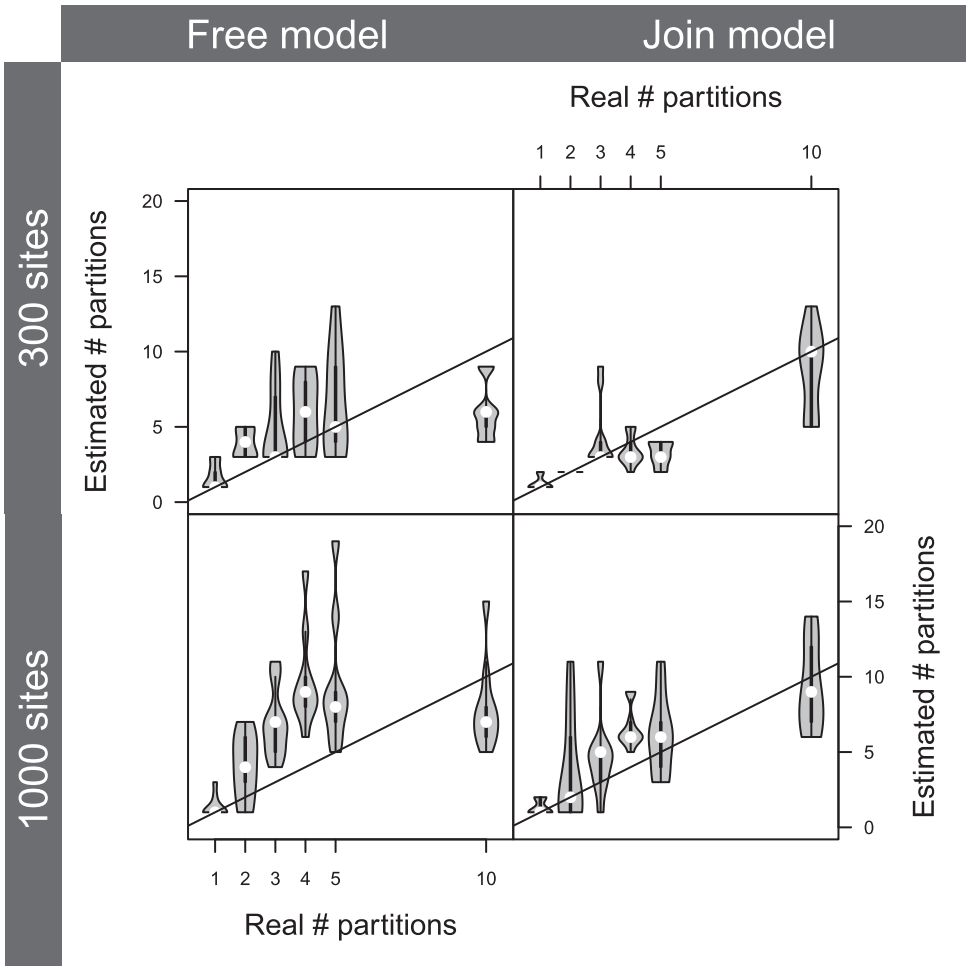

FIG. 2. Violin plots of the number of subsets retained under the BIC criterion as a function of the number of subsets used in the simulation procedure. Simulations with 1, 2, 3, 4, 5, and 10 random partitions were pooled (see Materials and Methods). The lines display the $y = x$ identity function.

overestimation of the number of clusters when the real number is small. As BIC is the most conservative criterion for model selection, this indicates that more permissive criteria like AIC or likelihood ratio test are likely to result in overparametrized models.

If there is an underlying discrete structure in the data, as in these simulations, the number of clusters inferred by our methods can provide insights into the biology of the organisms. In cases where the underlying structure is not discrete, but continuous, the partition itself is less relevant to the biology of the clade under study but should still be useful to decrease the noise in estimates of branch-wise parameters.

Heterogeneity in GC Content

The tree-of-life rRNA data set has been previously used to test branch-heterogeneous models in which the equilibrium GC content can vary among branches (Boussau and Gouy 2006; Dutheil and Boussau 2008). In prokaryotes, GC content in the stem portion of rRNA is correlated to optimal growth temperature (Galtier and Lobry 1997). Reconstructing GC content evolution therefore provides insight into phenotype evolution (Galtier et al. 1999; Boussau et al. 2008). The large number of leaves and relatively small number of complete sites in this data set make it challenging to estimate parameter-rich branch-heterogeneous models. Nonetheless, both the free and the join approaches recover extremely similar patterns of equilibrium GC content evolution along this tree (fig. 3). However, the free approach detects more changes between equilibrium GCs in clades, despite a lower number of clusters than the join approach (5 clusters for the free approach and 14 for the join). In addition, the join approach recovers more extreme values than the free approach. The differences between the two approaches in the branches leading from the root of the tree to its descendants are probably linked to different branch length estimates at the root. Such differences at the root between the two approaches are expected as it is known that it is very difficult to find the root of a phylogenetic tree using a branch-heterogeneous model of sequence evolution (Huelsenbeck et al. 2002; Boussau and Gouy 2006). Importantly, branches leading to thermophilic and hyperthermophilic species of bacteria and archaea, which live at high temperatures and have high GC contents in the stem portions of their rRNAs (Galtier and

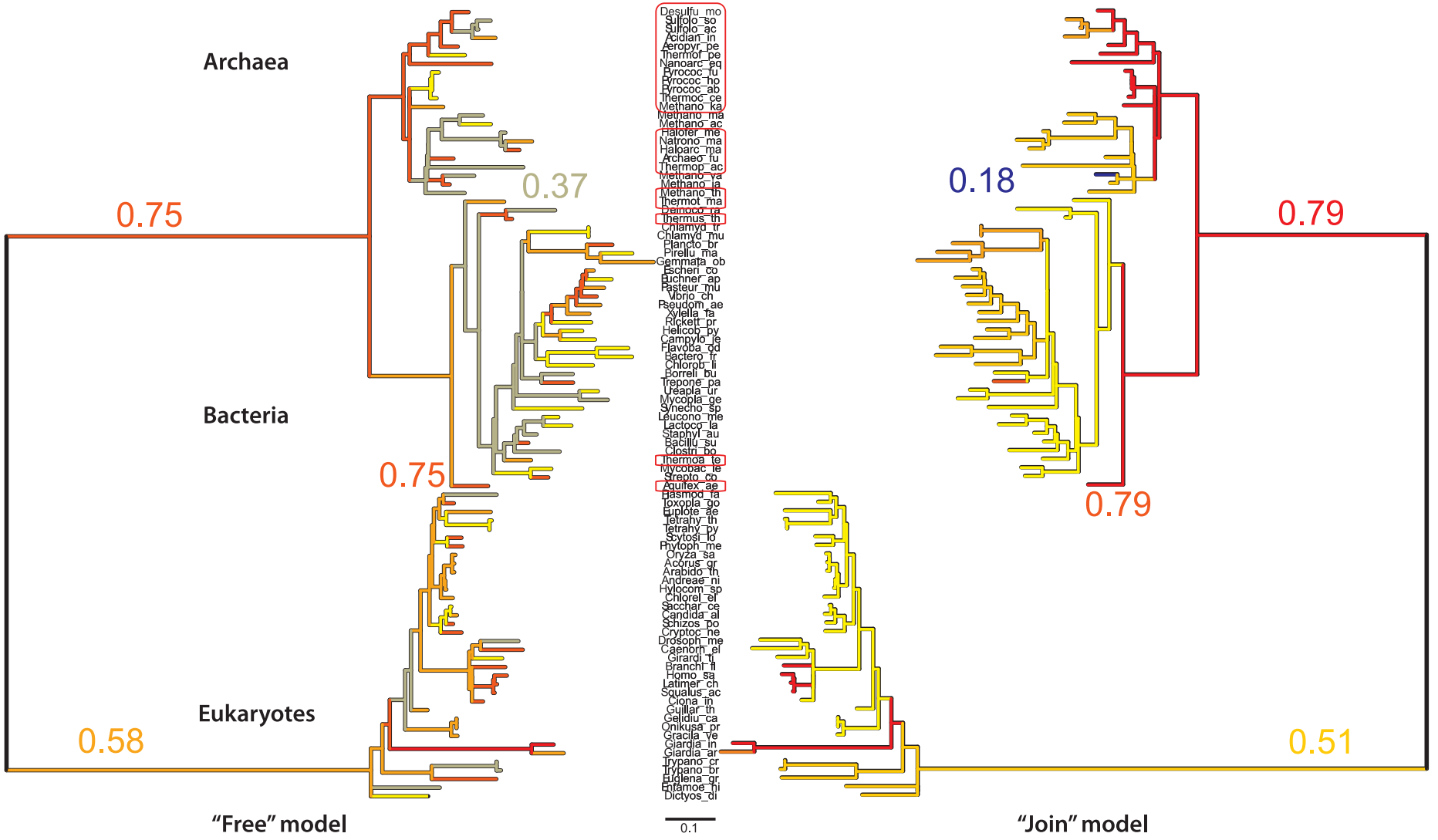

FIG. 3. Branch partition on the rRNA data set. Branches have been colored according to the equilibrium GC of the cluster they belong to (from blue for low GC to red for high GC, some values of interest are shown next to branches). Left: free model. Right: join model. Names of the thermophilic and hyperthermophilic species are circled in red. Branch lengths are proportional to the inferred numbers of substitutions per site (scale bar included).

Lobry 1997), are associated to high-equilibrium GC contents. The method also recovers high GC contents at the base of archaea and bacteria (Boussau et al. 2008) but does not seem to find evidence for later decreases in bacteria (Gaucher et al. 2008) and Archaea (Groussin and Gouy 2011), perhaps because the taxonomic sampling is inadequate to tackle such questions.

Expectedly, models obtained by our clustering methods have a much better BIC than a model (named 'general' here and in Dutheil and Boussau 2008) in which a specific stationary GC content parameter is associated to every single branch: the optimum free model has a log-likelihood of $-13,941$ and a BIC of 29,074 (182 branch lengths + transition/transversion ratio [κ] + shape of gamma distribution of site-specific rate [α] + GC content at the root node [θ_{root}] + 5 partition-specific equilibrium GC contents = 190 parameters), the join model a log-likelihood of $-13,956$ and a BIC of 29,159 (199 parameters, 14 clusters of branches), whereas the general model a log-likelihood of $-13,821$ and a BIC of 29,942 (182 branch lengths + κ + α + θ_{root} + 182 branch-specific equilibrium GC contents = 367 parameters) (Dutheil and Boussau 2008). This suggests that our clustering methods indeed find models with a good balance between parameter richness and fit to the data. Because we estimated 20 more complex models than the optimum one to ensure that the optimum had really been found, 23 (free approach) and 27 (join approach) models have been fully optimized and compared using BIC during model selection. This is far smaller than the total number of models possible for this large tree and even much smaller than the number of branches in the tree (181). Figure 4 shows the optimization profile with the values of equilibrium GC contents (θ) during optimization of increasingly complex models. Expectedly, models chosen according to AIC use more parameters than models chosen according to BIC. For each criterion, the global minimum is reached after local minima, which stresses the need for our algorithm not to stop at the first minimum value encountered. Despite this, our algorithms are efficient and end in less than an hour on a desktop computer for a data set containing 92 sequences.

Heterogeneity in Selection: Selecting Codon Models

Sex and dN/dS *in Daphnia*

This data set of 28 mitochondrial sequences from *Daphnia pulex* contains 14 sexual and 14 asexual strains (respectively noted *Sn* and *An*, $n \in [1:14]$) and has been used to study the impact of recombination between mitochondrial genome and nuclear genome on coding sequence evolution (Paland and Lynch 2006). Asexual strains are thought to have repeatedly evolved from sexual ancestors. Paland and Lynch (2006) computed dN/dS separately for sexual and asexual lineages and found it was higher in asexual lineages, in agreement with the expectation that recombination should improve the efficiency of purifying selection. Figure 5 displays the optimization profiles with the corresponding values of ω during optimization, and figure 6 shows that the free approach recovers the dN/dS difference between sexual and asexual lineages. It clusters branches in three groups with dN/dS values of 0.10, 0.25, and 0.85. Among terminal branches, for which the sexual/asexual status of the strain can be observed, all four branches assigned to the cluster with the largest dN/dS lead to asexual organisms, whereas only three branches leading to asexual organisms are in the cluster with the lowest dN/dS. Among the branches leading to sexual organisms, none are assigned to the cluster with the highest dN/dS but 11 are assigned to the cluster with the lowest dN/dS. Internal branches tend to be assigned to the cluster with the lowest dN/dS, which confirms that asexual strains originated from sexual ancestors. The join approach groups all branches but one in a single low dN/dS cluster. One branch stands out in a high dN/dS cluster of its own, the branch leading to the asexual strain A13. This is an indication that the join approach here is not appropriate as it aims at grouping together neighboring branches, when sampling was intentionally designed such that sister taxa have contrasted life-history traits. This shows that the two methods presented here should be used in situations where their respective underlying hypotheses fit the data set under study.

Breeding System and dN/dS *in Mantellid Frogs*

Mantellid frogs of Madagascar have been used to study patterns of ecology-driven diversification (Vences et al. 2002) as well as patterns of mitochondrial genome evolution (Kurabayashi et al. 2008). Ecologically, some species of frogs breed in ponds, whereas others breed in streams. This difference may have left a trace in the pattern of diversification of a particular genus of Mantellid frogs, *Boophis*, where pond breeders tend to speciate less easily than stream breeders, presumably due to lower barriers to gene flow in pond breeders (Vences et al. 2002). Recently, the data set analyzed in the present article, and initially published in Kurabayashi et al. (2008), was used to find that increases in mitochondrial genome sizes occurred jointly with increases in dN/dS (Boussau et al. 2011). This indicates that important changes in genome structure may have fixed non adaptively.

We used our approach to cluster branches showing similar dN/dS along the tree of 17 Mantellid frogs. Figure 7 shows the best partition obtained with the free and join approach using BIC. Both partitions are highly similar with two and five clusters, respectively, and show a striking difference in dN/dS between stream breeders (circled) and pond breeders. Overall, pond breeders tend to display lower dN/dS (about 0.045) than stream breeders (0.08). This dN/dS difference can be explained by selectionist or neutralist hypotheses. We know of no reason to assume different selection regimes in the mitochondrial genomes of stream breeders compared with pond breeders. Instead, because pond breeders may suffer less barriers to gene flow, they may have larger effective population sizes than stream breeders. With larger effective population sizes, purifying selection would be more efficient in pond breeders than in stream breeders, and their dN/dS would be lower.

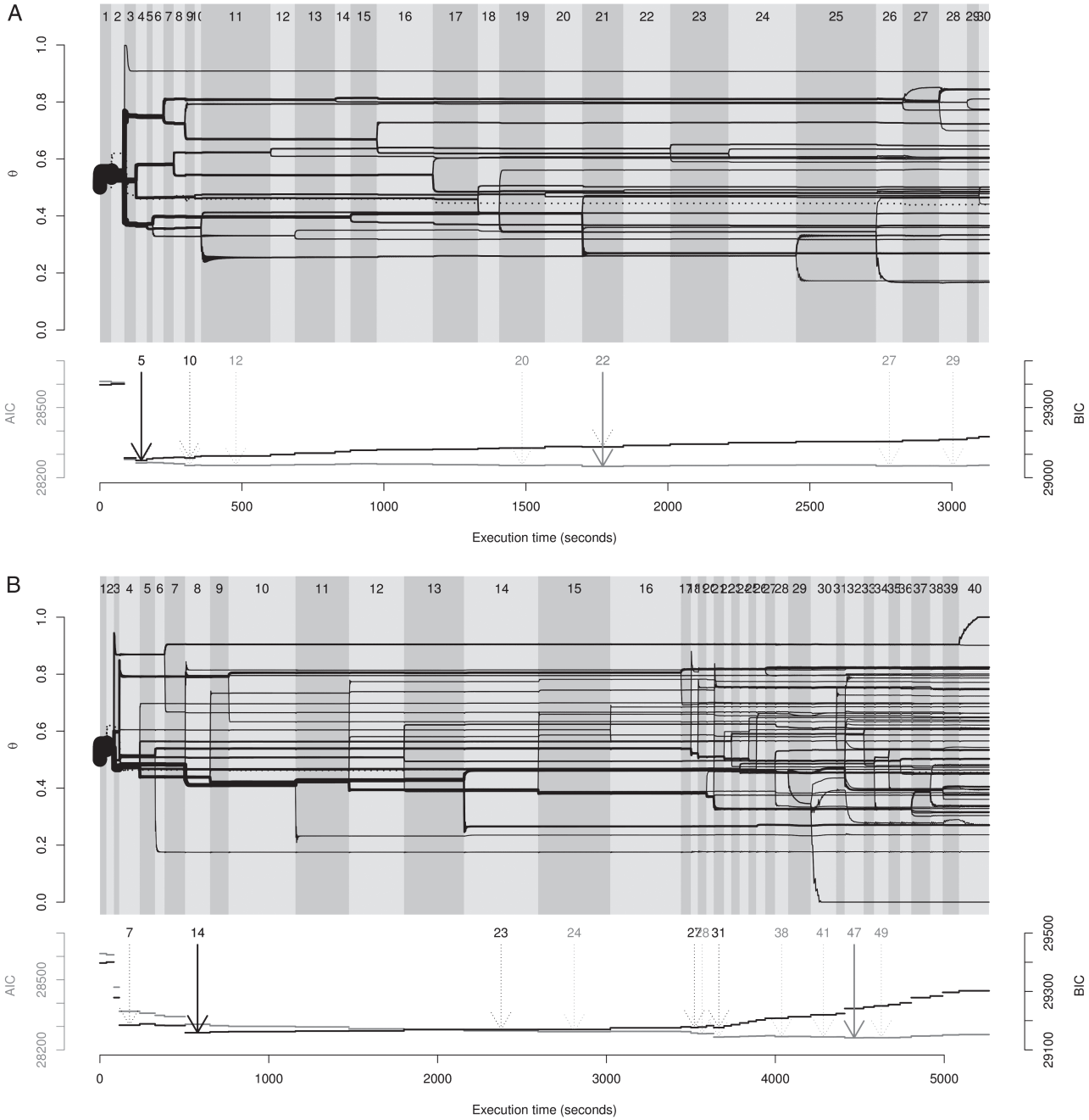

FIG. 4. Equilibrium GC content (θ_i parameters) estimation profiles for the rRNA data set. The values taken by the θ parameters during optimization (top panels) and the model comparison criteria (bottom panels) are plotted as a function of execution time in seconds, for both the free model (A) and the join model (B). Each distinct model is reported (top line numbers), from the simplest one (model 1, homogeneous, one partition with one θ on the left) to the most complex (models 30 and 40, more than 20 partitions/θ on the right). Top panels: line width depicts the size of the underlying partition. The dashed line shows the GC content at the root node of the tree. Bottom panels: AIC (gray) and BIC (black) values of corresponding models. Arrows depict local minima values, global minima being displayed with solid lines. Values on top of each arrow show the number of clusters in the corresponding model.

The optimal clusters found using BIC do not distinguish between branches with and without increases in mitochondrial sizes. This does not invalidate the results of Boussau et al. (2011) but instead illustrates the respective strengths of two different types of approaches. Hypothesis-driven approaches, as in Boussau et al. (2011), are very sensitive and therefore can reveal weak but significant effects. Exploratory approaches as used here cannot detect very subtle signals in the data but have the power to reveal strong signal of unsuspected but meaningful patterns of molecular evolution.

Innermost branches in the phylogenetic tree of Mantellids seem to be generally associated to larger values of dN/dS, both in the free and in the join approaches. This

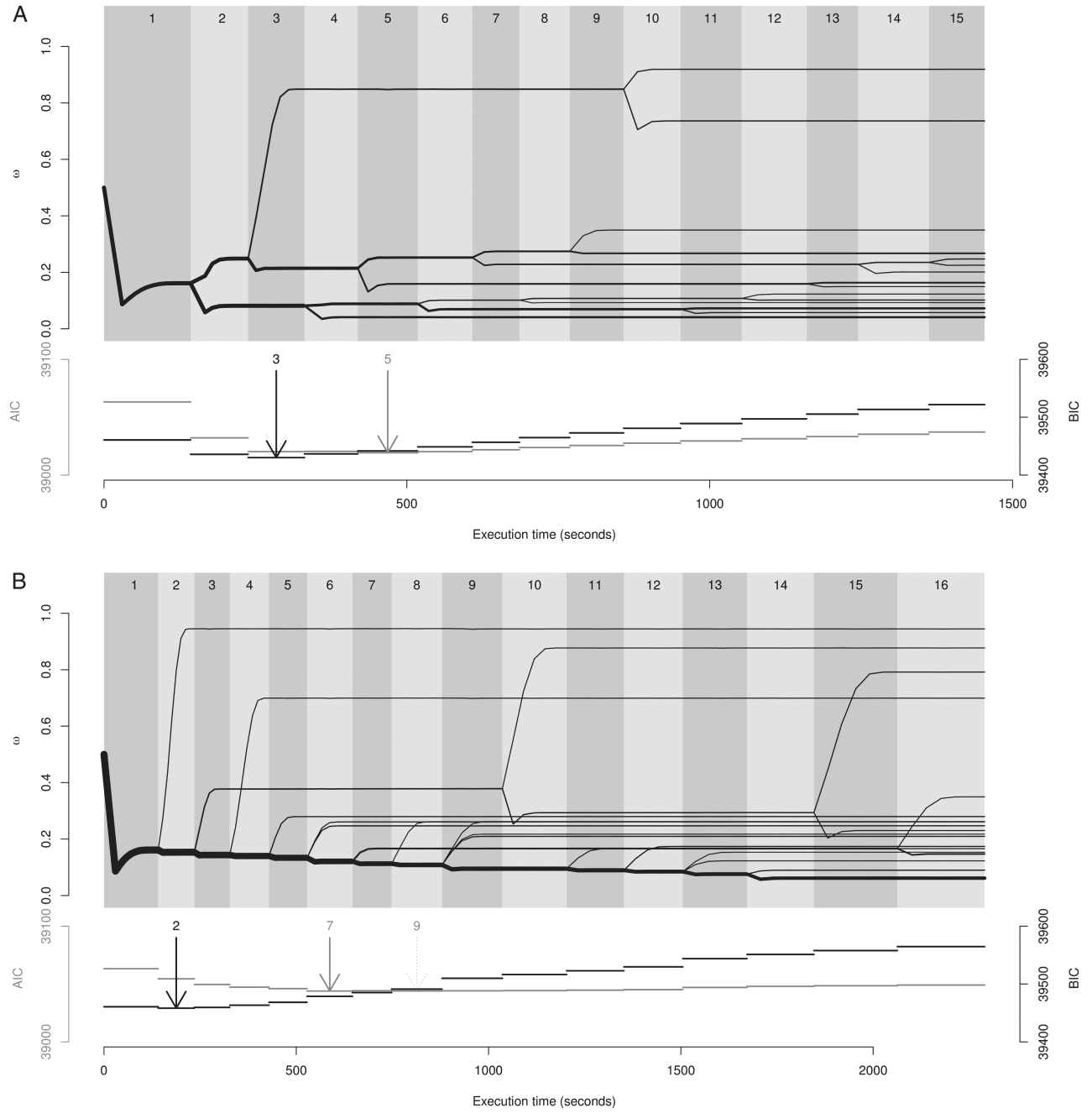

FIG. 5. dN/dS (ω_i parameters) estimation profiles for the Daphnia data set. The values taken by the ω parameters during optimization (top panels) and the model comparison criteria (bottom panels) are plotted as a function of execution time in seconds, for both the free model (A) and the join model (B). Other legends are similar to figure 4.

may be due to smaller effective population sizes in the ancestors of Mantellid frogs, about 20–50 Ma. This may also be indicative that dS is saturated on the most ancient branches, artificially inflating dN/dS. This second hypothesis would reconcile our results with the idea that Mantellid ancestors were pond breeders (Vences et al. 2002).

Assessing the Robustness of Selected Models

In the procedure we present, a homogeneous model is first used in order to perform substitution mapping, which is in turn used to cluster branches of the tree and guide the model selection procedure. The underlying rationale of this approach is the robustness of the substitution mapping procedure to the substitution model used (Minin and Suchard 2008). This robustness can be further assessed by performing a new substitution map from the selected model. This new a posteriori map can be used to obtain new branch-clustering trees and subsequent model selection. The resulting model can then be compared with the one previously found.

For codon data sets (Daphnia and Mantellid frogs), we find no difference in the selected models (with BIC criterion), with the exception of one node in the Daphnia data

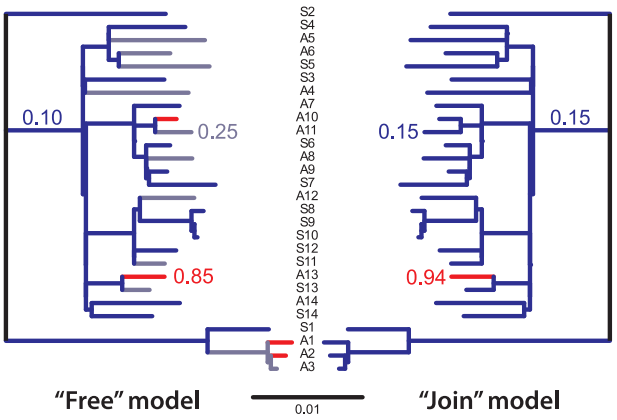

FIG. 6. Best partition proposed by the free and join models and according to BIC for the Daphnia data set. Strains named *An* are asexual, and strains named *Sn* are sexual. Branch lengths are proportional to the inferred numbers of substitutions per codon site (scale bar included).

set with a free clustering, moving from a partition with $\omega = 0.08$ to a partition with $\omega = 0.21$. For the rRNA data set, the free model gains one extra partition and the join model loses two. There are several differences between the two selected models in both cases, yet when results are compared on a per-node basis, resulting estimates of the θ parameters are quite similar (fig. 8). Varying nodes only move to the nearest cluster, reflecting uncertainty in the underlying value for the parameter, possibly due to a lack of signal in the data.

Discussion

We have presented an approach to cluster branches of a phylogenetic tree according to their pattern of sequence evolution. This approach provides models that accurately describe the evolution of a data set without overparametrization and bypasses preconception from the user. It can help in discovering new links between phenotype and sequence evolution.

Simulations and the four data sets we analyzed show that our method accurately clusters branches of a phylogenetic tree according to patterns of sequence evolution. This result represents a noticeable achievement as the challenge of finding an optimal partition of the branches of a tree is made difficult by the large combinatorial space that needs to be explored. An unrooted phylogenetic tree with 10 sequences contains 17 branches to cluster in 1–17 clusters, which can be done in more than 682 billion different ways (Zhang et al. 2011). Obviously all possible partitions cannot be tested, and heuristic algorithms have to be used. Such heuristics should only test a minute portion of the space of possible partitions to be fast but should not miss relevant partitions to provide biological insights.

Two recent works have shown that this challenge is currently generating a lot of interest and can now be tackled thanks to progresses in computing power (Zhang et al. 2011; Jayaswal et al. 2011). Zhang et al. (2011) developed a wrapper Perl script to produce option files to run PAML and test various partitions of branches in order to find branches showing similar dN/dS. Their most accurate algorithm starts by building a cluster around the single branch whose dN/dS is most different from the dN/dS of other branches. To find this branch, all models in which a branch is set apart from all other branches have to be optimized using PAML. The algorithm then clusters branches one at a time. Both the initial step and the recursive step can be long when the number of branches is large, and overall $O(n^2)$ models are optimized, with n the number of branches. We applied our methods to the lysozyme data set, originally studied by Yang (1998) and reanalyzed by Zhang et al. (2011). Using AIC, the free approach recovers their partition of the branches (not shown). Using BIC yields a lower number of clusters. However, branch-wise dN/dS values estimated using either AIC or BIC, and using either the free or join approaches, are robust and consistent with previous results (Yang 1998). Jayaswal et al. (2011) developed a heuristic algorithm in R to cluster branches based on the pattern of nucleotide substitutions rather than based on dN/dS. They use another type of algorithm, which starts from the most complex model in which each branch of the phylogenetic tree is associated to its own substitution matrix. Then, they iteratively cluster short branches or branches that have similar substitution matrices. In the end,

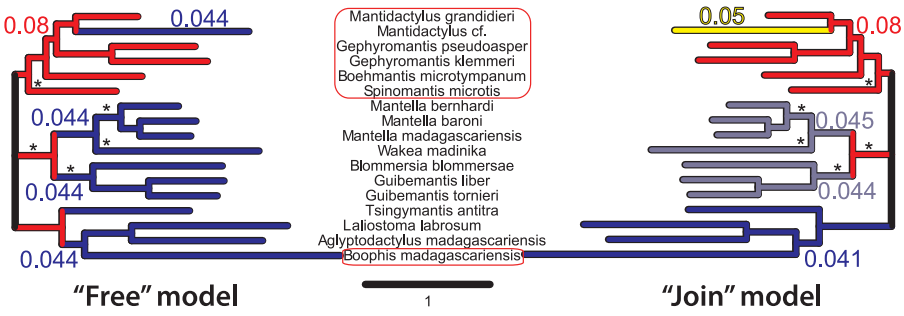

FIG. 7. Branch partition on the Mantellid frog data set. Branches have been colored according to the dN/dS of the cluster they belong to. dN/dS values are shown next to branches. Left: free model. Right: join model. Names of the species breeding in streams are circled in red; other species breed in ponds. Branch lengths are proportional to the inferred numbers of substitutions per codon site (scale bar included). Lineages where duplications have occurred are annotated with asterisks. Three of the five clusters recovered by the join approach have dN/dS values that are very close to each other: 0.041, 0.044, 0.045. The same branches in the free model are all found in the same cluster with a dN/dS of 0.044.

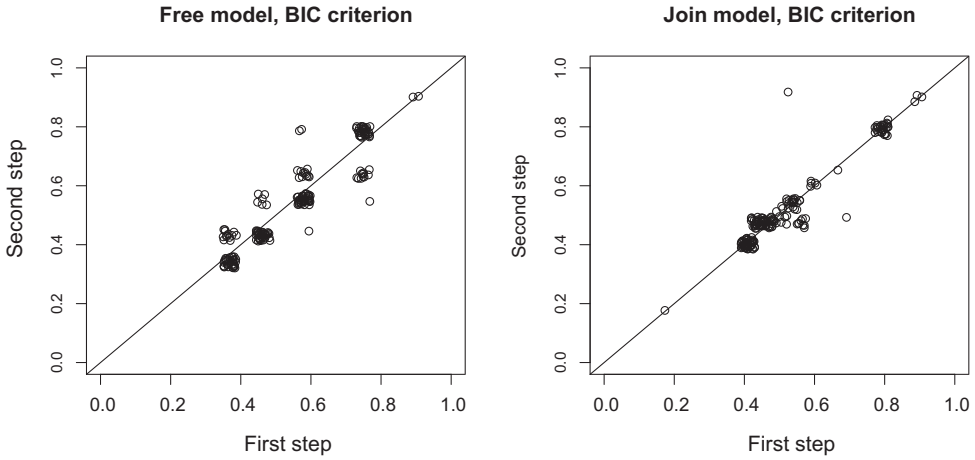

FIG. 8. Branch-specific equilibrium GC content estimates (θ parameters) from the best model according to BIC, after one step (x axis) or two steps (y axis) of model selection. Jitter was added in order to display overlapping points and do not correspond to real dispersion.

models ranging from the most complex, with one substitution matrix per branch, to the simplest, with one substitution matrix for all branches, are available, and the model providing the highest AIC or BIC is returned. This algorithm may be faster than Zhang et al.'s method as it requires optimizing only $O(n)$ models. However, the initial optimization of the parameters of the most complex model, the very same type of model discouraged in PAML's manual, may be difficult and costly. Both Zhang et al. (2011) and Jayaswal et al. (2011) approaches rely on a greedy heuristic algorithm to cluster branches, which cannot correct a mistake made at an early step. Therefore, if the initial decision to build clusters starting from the most distinct branch is not correct, or the estimation of substitution matrices in the most complex model is faulty, the resulting partition may not be correct.

Our approach improves upon these two early attempts in several respects. First, we cluster branches of the phylogenetic tree based on the substitution mapping procedure (Nielsen 2002; Minin and Suchard 2008), which has the advantage of being fast and robust to model misspecification (Minin and Suchard 2008). Substitution mapping provides counts of each type of substitution for each branch of the tree, based on a substitution model. Minin and Suchard (2008) have shown that substitution mapping was robust to the model used for mapping: even a simple homogeneous substitution model, where all branches share the same substitution matrix, can uncover accurate patterns of heterogeneity in the substitution process. Such a homogeneous substitution model also offers the advantage of being fast to fit, compared with parameter-rich, nonhomogeneous models used in the initial steps of the two previously mentioned approaches. The robustness of the mapping and subsequent clustering approach can be assessed a posteriori using the selected model in order to generate a new substitution map. In the data sets exemplified in this work, we showed that the resulting parameter estimates are quite robust to the initial model used for mapping substitution. However, it remains possible that for some data sets parameter estimates may be sensitive to the initial model. In such cases, successive iterations of mapping and model selection can be used to assess the uncertainty in branch-specific model attributions. Such an iterative approach can easily be performed with the TestNH package.

The robustness of the mapping procedure ensures that only meaningful clusters of branches are tested. This represents an advantage over the arbitrary starting point used by Zhang et al. (2011) and the complex and difficult to optimize starting point used by Jayaswal et al. (2011). Our procedure also prevents issues of overparametrization as it starts from models with small numbers of clusters. Counts of substitutions obtained for each branch are then used to partition branches in increasing numbers of subsets. From the resulting clustering tree, we design and fit a series of nested nonhomogeneous models, starting from the most simple one and progressively increasing the number of parameters, until it seems certain that improvement in AIC or BIC score can no longer be achieved. This procedure has the advantage that less than n optimizations are required (45 or 49 optimizations for the rRNA data set, containing 181 branches) and, perhaps more importantly, the models with the largest numbers of clusters, overparameterized and most costly to optimize, are often never optimized as the algorithm settles on models with small numbers of clusters (for instance, two to five dN/dS clusters in the tree of Mantellid frogs, which contains 31 branches). This is in sharp contrast to Jayaswal et al. (2011)'s approach where all the most parameter-rich models are necessarily optimized.

We propose in this work two clustering algorithms, corresponding to two distinct assumptions about character evolution. Choosing between these two methods mostly depends on the biological question underlying each specific analysis. If our approach is to be used mainly for avoiding overparametrization issues while estimating branch-specific parameters, we recommend using the free algorithm, as we expect it to provide better models (according to statistical criteria like BIC).

Another important feature of our algorithms is their generality as they do not depend on a particular sequence type or family of models. They are by construction not limited to d*N*/d*S* studies (Zhang et al. 2011) or to studies of compositional heterogeneity in DNA sequences (Jayaswal et al. 2011): built upon the Bio++ libraries (Dutheil et al. 2006), these programs can cluster branches based on any features of DNA, RNA, codon, or amino acid substitution matrices.

Finally, as shown in the simulations and on real data, these algorithms are fast despite the number of models fitted. This stems from our use of C++ for all steps of the code and of substitution mapping to restrain the number of models to optimize. Even with conservative parameters allowing to test a large set of models, the total execution times for the data sets exemplified in this work are of the order of 1–3 h on a 2.27-Ghz computer, which is comparable to the execution time of PAML with a branch model. Our results suggest that less conservative parameters for model exploration can be safely used: in our examples looking at only three models after a local AIC or BIC minimum has been found would still ensure that the global minimum is reached and would provide a significant gain in execution time.

We are confident that our two clustering approaches will be very useful to study patterns of molecular evolution since they can be used to cluster branches according to any statistics describing sequence evolution. For instance, one could cluster branches according to their ratio of transitions to transversions, according to both their equilibrium GC content and their d*N*/d*S*, or according to counts of all possible types of substitutions. Moreover, our programs can run on DNA, RNA, codon, or amino acid sequences, are efficient, and require very little input from the user. They could therefore be used on phylogenomic data sets to estimate genome-wide heterogeneity in sequence evolution, help reveal phenotypic determinants of sequence evolution, and consequently provide means to reconstruct phenotypic evolution along the tree of life.

Supplementary Material

Supplementary materials are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors would like to thank Asger Hobolth and Paula Tataru for sharing code on the computation of the conditional number of jumps of a Markov chain. We greatly benefited from some C++ code contributed by Laurent Guéguen. We would also like to thank two anonymous reviewers for very constructive comments on an earlier version of this manuscript. This work was supported by European Research Council grant 232971 ("PopPhyl") and the French Agence Nationale de la Recherche "Bioinformatique" (ANR-10-BINF-01 "Ancestrome"). B.B. was supported by a postdoctoral fellowship from the Human Frontier Science Program and the CNRS. This publication is contribution 2012-016 of the Institut des Sciences de l'Évolution de Montpellier (UMR 5554CNRS).

References

Aris-Brosou S. 2007. Dating phylogenies with hybrid local molecular clocks. *PLoS One* 2(9):e879.

Boussau B, Blanquart S, Necsulea A, Lartillot N, Gouy M. 2008. Parallel adaptations to high temperatures in the Archaean eon. *Nature* 456(7224):942–945.

Boussau B, Brown JM, Fujita MK. 2011. Nonadaptive evolution of mitochondrial genome size. *Evolution* 65(9):2706–2711.

Boussau B, Daubin V. 2010. Genomes as documents of evolutionary history. *Trends Ecol Evol.* 25(4):224–232.

Boussau B, Gouy M. 2006. Efficient likelihood computations with nonreversible models of evolution. *Syst Biol.* 55(5):756–768.

Bromham L. 2009. Why do species vary in their rate of molecular evolution? *Biol Lett.* 5(3):401–404.

Drummond AJ, Suchard MA. 2010. Bayesian random local clocks, or one rate to rule them all. *BMC Biol.* 8:114.

Dutheil J, Boussau B. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evol Biol.* 8:255.

Dutheil J, Gaillard S, Bazin E, Glémin S, Ranwez V, Galtier N, Belkhir K. 2006. Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinform.* 7:188.

Dutheil J, Pupko T, Jean-Marie A, Galtier N. 2005. A model-based approach for detecting coevolving positions in a molecule. *Mol Biol Evol.* 22(9):1919–1928.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform.* 5:113.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol.* 17(6):368–376.

Galtier N, Boursot P. 2000. A new method for locating changes in a tree reveals distinct nucleotide polymorphism vs. divergence patterns in mouse mitochondrial control region. *J Mol Evol.* 50(3):224–231.

Galtier N, Gouy M. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol Biol Evol.* 15(7):871–879.

Galtier N, Lobry JR. 1997. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J Mol Evol.* 44(6):632–636.

Galtier N, Tourasse N, Gouy M. 1999. A nonhyperthermophilic common ancestor to extant life forms. *Science* 283(5399):220–221.

Gaucher EA, Govindarajan S, Ganesh OK. 2008. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* 451(7179):704–707.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol.* 27(2):221–224.

Groussin M, Gouy M. 2011. Adaptation to environmental temperature is a major determinant of molecular evolutionary rates in Archaea. *Mol Biol Evol.* 28(9):2661–2674.

Heath TA, Holder MT, Huelsenbeck JP. Forthcoming 2011. A Dirichlet process prior for estimating lineage-specific substitution rates. *Mol Biol Evol.*

Hickey DA, Singer GA. 2004. Genomic and proteomic adaptations to growth at high temperature. *Genome Biol.* 5(10):117.

Hobolth A, Stone EA. 2009. Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution. *Ann Appl Stat.* 3:1204.

Huelsenbeck JP, Bollback JP, Levine AM. 2002. Inferring the root of a phylogenetic tree. *Syst Biol.* 51(1):32–43.

Jayaswal V, Ababneh F, Jermiin LS, Robinson J. 2011. Reducing model complexity of the general Markov model of evolution. *Mol Biol Evol.* 28(11):3045–3059.

Kurabayashi A, Sumida M, Yonekawa H, Glaw F, Vences M, Hasegawa M. 2008. Phylogeny, recombination, and mechanisms of stepwise mitochondrial genome reorganization in mantellid frogs from Madagascar. *Mol Biol Evol.* 25(5):874–891.

Lartillot N, Poujol R. 2011. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Mol Biol Evol.* 28(1):729–744.

Minin VN, Suchard MA. 2008. Fast, accurate and simulation-free stochastic mapping. *Philos. Trans. R. Soc. B* 363(1512):3985–3995.

Nielsen R. 2002. Mapping mutations on phylogenies. *Syst Biol.* 51(5):729–739.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148(3):929–936.

Paland S, Lynch M. 2006. Transitions to asexuality result in excess amino acid substitutions. *Science* 311(5763):990–992.

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20(2):289–290.

Pupko T, Sharan R, Hasegawa M, Shamir R, Graur D. 2003. Detecting excess radical replacements in phylogenetic trees. *Gene* 319:127–135.

Rodrigue N, Philippe H, Lartillot N. 2008. Uniformization for sampling realizations of Markov processes: applications to Bayesian implementations of codon substitution models. *Bioinformatics* 24(1):56–62.

Sainudiin R, Wong WSW, Yogeeswaran K, Nasrallah JB, Yang Z, Nielsen R. 2005. Detecting site-specific physicochemical selective pressures: applications to the Class I HLA of the human major histocompatibility complex and the SRK of the plant sporophytic self-incompatibility system. *J Mol Evol.* 60(3):315–326.

Tamura K. 1992. The rate and pattern of nucleotide substitution in *Drosophila* mitochondrial DNA. *Mol Biol Evol.* 9:814–825.

Tataru P, Hobolth A. 2011. Comparison of methods for calculating conditional expectations of sufficient statistics for continuous time Markov chains. *BMC Bioinform.* 12(1):465.

Vences M, Andreone F, Glaw F, Kosuch J, Meyer A, Schaefer HC, Veith M. 2002. Exploring the potential of life-history key innovation: brook breeding in the radiation of the Malagasy treefrog genus *Boophis*. *Mol Ecol.* 11(8):1453–1463.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol.* 15(5):568–573.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24(8):1586–1591.

Yang Z, Roberts D. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. *Mol Biol Evol.* 12(3): 451–458.

Yang Z, Yoder AD. 2003. Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene Loci and calibration points, with application to a radiation of cute-looking mouse lemur species. *Syst Biol.* 52(5):705–716.

Zhang C, Wang J, Xie W, Zhou G, Long M, Zhang Q. 2011. Dynamic programming procedure for searching optimal models to estimate substitution rates based on the maximum-likelihood method. *Proc Natl Acad Sci U S A.* 108(19):7860–7865.